I J C T A, 9(17) 2016, pp. 8755-8763

Segmentation Free Approach for the Recognition of Hindi Compound Characters

Pratibha Singh[*] and Priyank Verma[**]

ABSTRACT

The recognition of Hindi Characters is of two types, one is simple characters recognition that comprises of consonants and vowels and second is compound characters recognition. The compound characters are those which are formed by joining of two or more consonants. So the recognition of compound characters is more difficult in comparison to that of simple characters due to their structure. This paper presents an approach for Hindi compound character recognition using various classifiers like QDC, LDC, KNNC, BPXNC, SVM and NMC. For computing the character's features we took complete characters along with Shirorekha and without segmentation of characters. We obtained the recognition rate 65.28% of Hindi Compound Characters by using LDC classifier.

Keywords: Compound Characters,Pre-processing, Feature extraction, Gradients.

I. INTRODUCTION

In today's world, the modern communication is done in the form of electronic documents as soft copies for sharing of information. These soft copies are the easy medium of sharing and storage of documents in real time environment and also secure and immediate way of communication. But still handwritten documents that are used as communications for office works and in other applications like automatic sorting of postal mail, bank cheques, car plate numbers, zip code and other office works. Recognition of Hindi handwritten character is one of the major problem due to their cursive structure [1]. Hence an automated tool is required to analyze these handwritten characters or documents. For the researchers, the recognition of handwritten characters of Indian scripts is a challenging task. The reason behind this is that the size of Hindi characters including with their complex shape and with different modifiers vary in accordance with different writers [2]. Hindi that is derived from Devanagari script, is a most used language of India and 15th most spoken language in the world.

Hindi handwritten characters are not recognized efficiently and accurately by computer machine. But typed Hindi characters can be easily recognized by computer machine [3]. People write Devanagari script in their unique ways with different types of pens or pencils with variant thickness. So it is very difficult to recognize this script due to the variant forms of writing styles. Hindi language is constituted using various symbols which is divided into 13 vowels and 36 consonants, which are shown in Figure 1.

As shown in figure 1 there are 13 vowels which are called in Hindi language as SWARS and 36 consonants which are called as VYANJANS. In English language we have alphabets as A, B, C, D and so on up to Z. There are the total 26 alphabets which are used to make the complete word and words to make the complete sentence. In a same manner in Hindi language the SWARS and VYANJANS as shown in figure 1 are used to make the words which are called 'SABD' and these words or SABD are used to make the complete sentence which is called 'VAKYA' in HINDI language. But there are also some modifiers used with these SWARS and VYANJANS which are shown in Figure 2.

[*] Assistant Professor, Electronics & Instrumentation Engg., IET DAVV, Indore, India, *E-mail: prat_ibh_a@yahoo.com*

[**] Student, M.E(Digital & Instrumentation Engg.), IET DAVV, Indore, India, *E-mail: priyanksati@gmail.com*

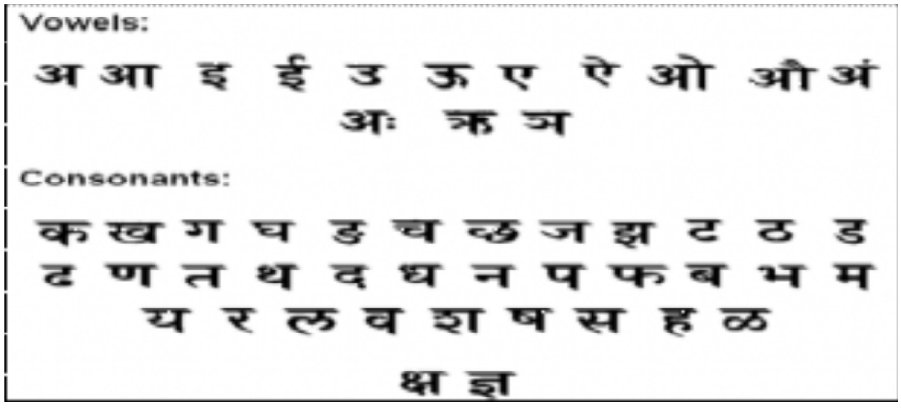

Figure 1: Hindi Language Font

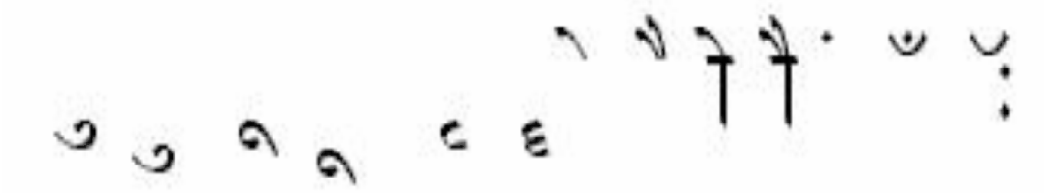

Figure 2: Modifiers of Hindi Language

There is a top line along which the words are written. In figure 3 the word HINDI is shown written in Devanagari script. In this modifiers are also used to make the complete word or SABD. The upper horizontal line is used to connect the letters is called 'SHIROREKHA'. Hindi is written from left to right. It is a cursive shape language in which the symbols are mostly having curves which is clear from Figure 3.

Figure 3: An Example of SABD

This writing style is divided into three sections or zones which are shown in below Figure 4.

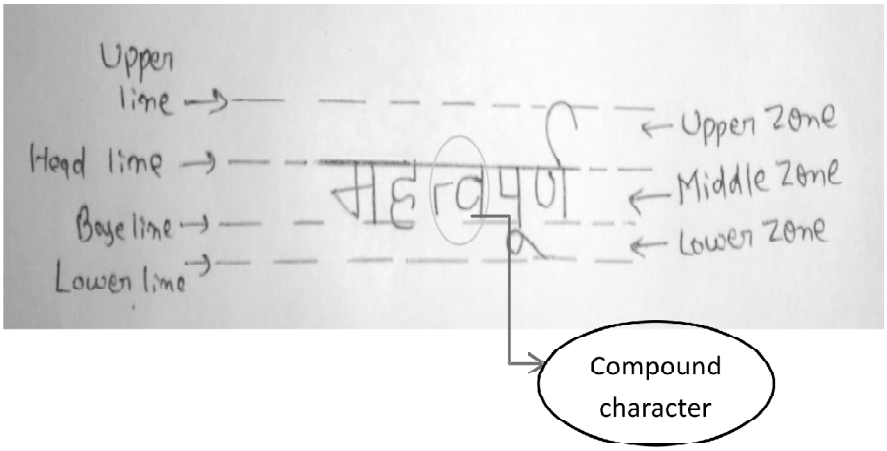

Figure 4: HINDI Language Pattern

There are 3 zones present in a word with Head line or 'SHIROREKHA': Upper zone, Middle zone and Lower zone. The modifiers are used in the upper zone and lower zone as and when required. In the word shown in figure 4, there is one compound character which is also called conjunct character. The conjunct character is formed with the combination of half and full letter of consonants. This paper is about the recognition of these isolated compound characters.

There are two types of character recognition approaches online character recognition and offline character recognition. These two categories are specifically used for hand written character recognition. The whole classification of character recognition methods is shown in Figure 5.

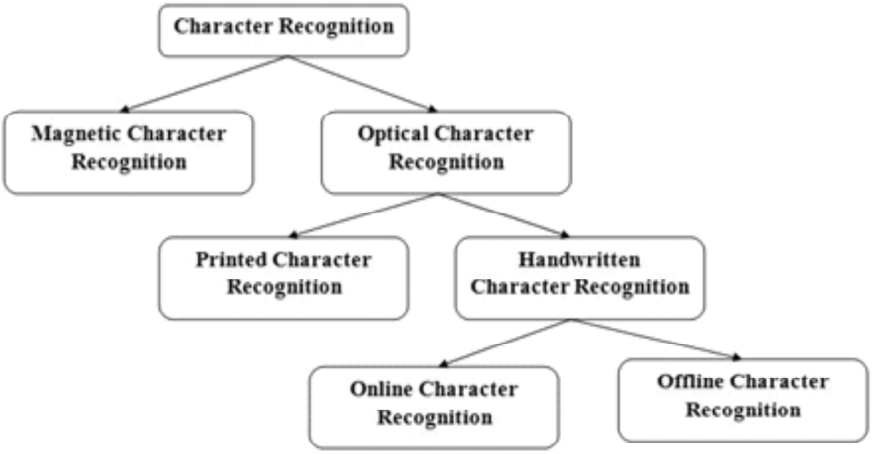

Figure 5: The Classification of Character Recognition Methods

Numerous research has been done in the field of character recognition. Namrata et al. [14] in 2011 presented work for handwritten Marathi character recognition. In this work, they used structured Marathi based feature extraction methods and found the recognition efficiency of words from 85% to 90%. They used polygon or box fitting approach for feature extraction and rule based approach for classification. Another work is done in this field by K.V. kale et al. [15] in 2014, by using Zernike moment based feature extraction method on compound characters of Marathi script. As Marathi compound characters have typical variations in their shapes, so they used Zernike moment technique which has a property of rotation invariance. They used 27000 handwritten character images and used zoning method after resizing and pre-processing them into 30×30 pixels sizes. Then they applied Zernike moments to each zone and found the recognition rate. The percentage recognition reported was 98.37% and 95.82% by SVM and KNN classifier respectively. Another work on Handwritten Marathi compound character recognition is done by Sushama Sheilke et al. [16] in 2011. In their work they used neural network as a classifier and found the recognition accuracy of 97.95%. They used modified wavelet approximation and Euclidean distance features from structurally classified and normalized characters. Md. Mahbubar et al. [6] in 2015 presented work for Bangla handwritten character recognition, they used convolutional neural network (CNN) in their work. The handwritten images were normalized and then CNN is used to classify individual characters. It does not employ any feature extraction method like others related works. 20000 handwritten characters with different shapes and variations were used in their study. They used distinct feature extraction techniques and various classification tools in their recognition schemes. Recently, Convolutional Neural Network (CNN) is found efficient for English handwritten character recognition. Other character recognition work [7-8] [4] is done on Sanskrit script and numeric data digit, in which SVM and neural network was used as classifiers after pre-processing and feature extraction methods by many researchers and they found the good recognition rate on large database of handwritten characters.

In this paper, we proposed a system for Hindi handwritten compound character recognition without separation of these compound characters. Some work for handwritten number digit system and handwritten Hindi of single characters and single Bangla, Urdu and Gurumukhi characters are found but no work for Hindi handwritten compound characters recognition is found so far to the best of our knowledge.

This paper discusses the various feature extraction methods for characters recognition. The rest of the paper is described as follows. Section 2 describes data collection, various pre-processing methods of character recognition, proposed methodology of feature extraction and various structural classification methods: using back propagation neural network, QDC, LDC, NMC,ANN and KNN. Section 3 presents experimental results using classifiers for recognition of Hindi compound characters. Section 4 describes the conclusion.

II. PROPOSED METHODOLOGY

Our proposed approach is to recognize Compound Characters of Hindi. For this we have created a dataset of 40 compound characters in ".jpeg" format written by 100 people from different backgrounds such as from college students, faculties, teachers, housewives etc. So we have 4000 compound characters. In which there are total 40 classes of each different characters. The various classes of compound characters used in the present study is shown in figure 6.

Figure 6: Sample Compound Character Data

This is a dataset of compound characters written by one person and like this there are total 100 tables written by different persons. We divide this table into 40 different classes and in each class there is same compound characters written by 100 people. These datasets are divided into Training and Testing sets for recognition purposes. There are different stages for handwritten HINDI character recognition like Pre-processing, feature extraction and classification as shown in Figure 7.

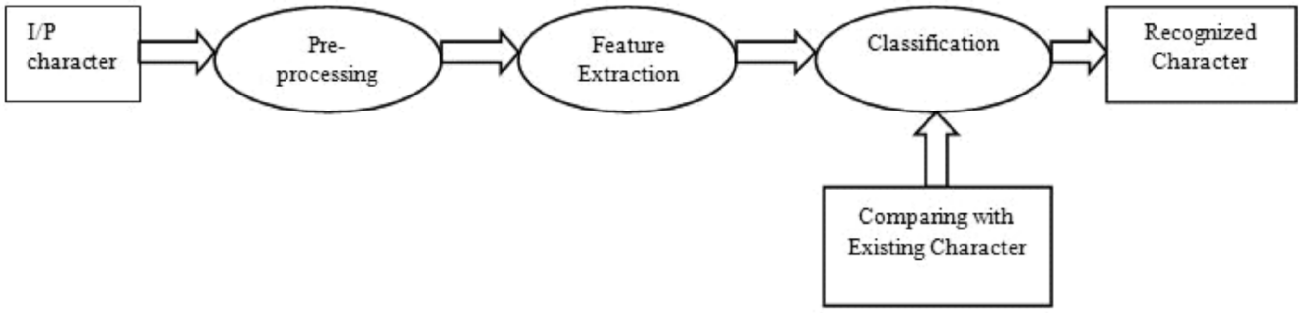

Figure 7: The Flow Algorithm for Character Recognition

2.1 Pre-processing Steps

Before Pre-processing, Image Acquisition process is performed in which input characters of Hindi script are scanned by the scanner in jpeg format of 300 dpi. Hence in image acquisition process the input is in handwritten character format and output is in digital format for pre-processing. Now for the improvement of the image quality the pre-processing is used in which Noise Reduction, Binarization, Morphological operations, Skeletonization, Normalization and Scaling of input image take place.

2.1.1 Noise Reduction

In image the noise can be because of poor scanning or due to the writing instrument which is unwanted and should be removed for further process. Some filtering methods are used for the noise reduction like Butterworth low pass filter, Median filter, Min-Max filter and so on.

2.1.2 Binarization

Binarization is the process of converting gray scale or colored image into binary form (0 & 1) by thresholding [9]. Methods of thresholding to convert image into binary are - Adaptive thresholding, Global thresholding and local thresholding.

2.1.3 Skeletonization

It is the process to reduce the input image of character into thin line format without changing the structure of the input image. It uses the 'bwmorph' function [10].

2.1.4 Normalization

In this the input character image is resized into the standard format through "imresize" command in MATLAB.

2.2 Feature Extraction

To define the shape of character, the Feature extraction of image is needed. To achieve the high performance of recognition of characters, the selection of feature extraction method is one of the most important thing. Feature Extraction methods can be classified into three categories as follows:Statistical Features, Geometric and topological Features and Global transformation and series expansion.

Out of many feature extraction methods this work is based on calculating directional feature over nine partitions for database of Hindi compound characters. The steps of feature extraction are elaborated as follows:

2.2.1 Zoning

Zoning is a statistical feature based algorithm which is used for reducing the dimension of feature set [11]. The 'Zoning' based feature extraction method in this research provides better result, even when slant removing, smoothing and filtering are not considered in the pre-processing steps. In this an 3×3 rectangular

grid is superimposed on our character image and for each 3×3 zones, the average gray level is computed. The density of each character image pixels is calculated of all overlapping or non-overlapping zones in the rectangular frame.

2.2.2 Eight Directional Feature Extraction (8-DGF)-

To find the gradients of an image, each image of Hindi character is normalized into 32×32 size. Sobel mask operator is used to calculate the horizontal gradient (Gx) and vertical gradient (Gy) components which are shown in Figure 8.

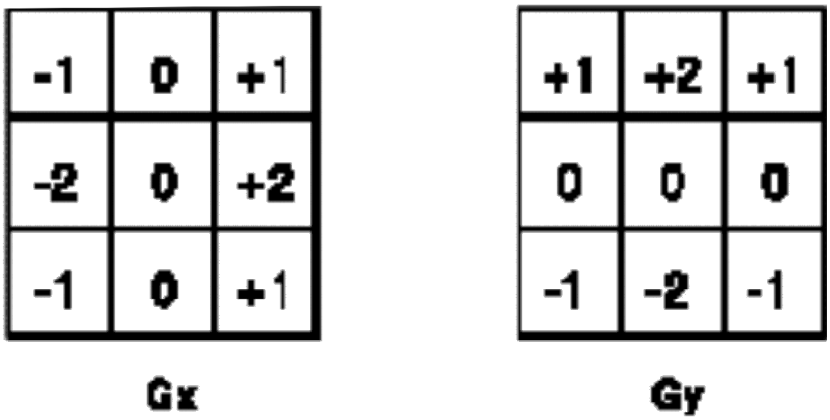

Figure 8: Sobel Operator

We calculate the gradient of a pixel (i,j) by using following formula.

$$Gy = gh(i, j) = f(i-1, j-1) + 2f(i-1, j) + f(i-1, j+1) - f(i+1, j-1) - 2f(i+1, j) - f(i+1, j+1) \quad (1)$$

$$Gy = gh(i, j) = f(i-1, j-1) + 2f(i-1, j) + f(i-1, j+1) - f(i+1, j-1) - 2f(i+1, j) - f(i+1, j+1)$$

$$grad = Gy/Gx = tan^{-1}[gh(i, j) / gv(i, j)] \quad (2)$$

It has been assumed that whenever a pixel is surrounded by all black pixels then its gradient values will be taken 1.

Gradients are divided into 8 equal parts by Sobel operator that depends on the values fall between the ranges of angles [13].

2.3 Classifiers

On the basis of training data classifier model is created. The model is given the testing data after which classification categories are defined by the classifier. The term "classifier" sometimes refers to the mathematical function, implemented by a classification algorithm that maps input data to a category. Some classifiers which are used in this work are as follows:

2.3.1 Linear Discriminate Classifier(LDC)

Linear Discriminate analysis (LDA) is a classification method originally developed in 1936 by R.A. Fisher. LDA is based upon the concept of searching for a linear combination of variables (predictors) that best separates two classes (targets)[18]. It is simple, mathematically robust and often produces models whose accuracy is as good as more complex methods. A linear classifier is a mapping which partitions feature space[17] using a linear function (a straight line, or a hyper plane).It separates two classes using a straight line in feature space. In two dimensions the decision boundary is a straight line. The LDC can be classified into categories as linear separable data and linear non-separable data. This is used for character recognition, face recognition, bankruptcy prediction, earth science and in biomedical studies.

2.3.2 Quadratic Discriminate Classifier (QDC)

A Quadratic Classifier is used in statistical classification and machine learning for the separate measurements of two or more classes of events or objects by a quadratic surface. Quadratic Discriminate Analysis (QDA) and Linear Discriminate Analysis (LDA) are closely related. In QDA there is no assumption that the covariance of each classes is identical as compare to LDA.

2.3.2 K- Nearest Neighbor classifier (KNN)

KNN method is used in pattern recognition for regression and classification and it is a non- parametric method. In both the cases of classification and regression, input consists of k closest training examples in the feature space. Its result depends on whether it is used for regression or classification. KNN is the simplest algorithm among the entire machine learning algorithm. KNN classifier is easy and works better for recognition of simple problems but the KNN is not good for noisy data and is slow learner.

*k-NN Classification***:** In this using Euclidian distance or any other distance method, the input test sample is classified to class membership that is based on training samples features matching.

k-NN Regression: To estimate continuous variables of k-NN algorithm, k-NN regression is used. In this the result belongs to the object property value. This value is computed from its nearest neighbors mean values.

A ‘matching matrix’ or ‘confusion matrix’ is used as a tool to validate accuracy of k-NN classification.

2.3.3 Support Vector Machine (SVM)

SVM are supervised learning models in machine learning with associated learning algorithms that analyze data use for regression and classification analysis. SVM is a two-class classifier, it categorize the classes by a decision boundary. An SVM model presents examples as point in space, mapped so that the examples of the separate categories are divided by a clear wider gap and new examples are mapped into same space and predicted to belong to a category based on which side of gap they fall on.

SVM can also perform non-linear classification in addition to performing linear classification using kernel trick, implicit mapping their inputs into high-dimensional feature spaces. SVMs are widely used in Hand-written character recognition, text and hypertext categorization, in Biomedical and other sciences. The main principal of an SVM is to map input data onto a higher dimensional feature space non linearly related to the input space and to determine separating hyper plane with the maximum margin between two classes in feature space.

2.3.4 Nearest Mean Classifier

Nearest Mean Classifier (NMC) with its similarity to Rocchio algorithm, is also called as Rocchio Classifier. Nearest Mean Classifier or Nearest Centroid Classifier, in machine learning, is used as a classification model which assigns the observations, the label of class of the training samples whose centroid or mean is very closed to the observation.

2.3.5 Artificial Neural Network

Neural networks is made up of layers. Layers consist of interconnected ‘nodes’ containing an activation function. Patterns are applied to network through ‘input layer’ which then communicates to one or more than one ‘hidden layer’. The weighted ‘connections’ are used for actual processing where the values of weights are modified as the training progresses. After hidden layers there is a final ‘output layer’ where the output of testing patterns are represented in the form of proximity.

An ANN is composed of many artificial neurons that are linked together according to a specific network architecture. The objective of the neural network is to transform the inputs into meaningful outputs. ANNs learn with the help of examples in the similar way as the child starts to learn to recognize cats from example of cats. ANN contains 'learning rule' to modify the weights of connections according to the input patterns which is presented with. Widely used 'learning rule' used by neural networks is the 'delta rule' utilized by most common class of ANNs called 'Back Propagation Neural Networks' (BPNNs). With delta rule, 'learning' is a supervised process which occurs with each cycle by a forward activation flow of outputs and backwards error propagation of weight adjustments. Back propagation adjusts the weights of the neural network in order to minimize the network's total mean squared error. The total mean squared error is error of output neuron k after the activation of the network on the nth training. In our research the ANN is used as a classifier to classify the different dataset after training and testing through neural network learning and simulation.

III. EXPERIMENTAL RESULTS

The training and testing patterns undergo pre-processing and feature extraction steps. In this work, firstly the bounding box is divided into 9 zones for each data image. Then gradients of each of 9 zones of 3×3 zoning is calculated for each data image. Categorization in classes is performed using various classification methods namely LDC, QDC, NMC, BPXNC, KNNC and SVC. The experimental results are as shown in Table 1 and Table 2. The error rate is calculated using k-fold cross validation and a fixed percentage of training data. Training percentage is varied and the performance is given in Table1 (on 80%, 70% & 60% data images). The obtained results using k-fold cross validation on different classifiers is given in Table 2. The error rate is calculated using k- fold cross validation and gradient features wherethe number of fold is taken to be 5 fold in cross validation.

Table 1
Error rate using holdout cross validation & gradient features

Training dataclassifiers	*80%*	*70%*	*60%*
BPXNC	40.75%	37.167%	95.1%
KNNC	46.125%	49.91%	48%
QDC	38.125%	37.08%	39%
LDC	30.00%	34.8%	33.87%
NMC	48.75%	47.58%	49.18%

Table 2
Error rate using k fold Cross validation & gradient features

Classifierk-folds	*SVC*	*BPXNC*	*KNNC*	*QDC*	*LDC*	*NMC*
1st	40.87	36.37	49.75	35.75	33.25	47.87
2nd	41.75	40.00	49	33.87	33.25	50.25
3rd	42.5	38.12	46.37	37.62	35.62	50.62
4th	42.87	40.12	47.37	35.75	36.25	50
5th	40.37	35.5	48.12	38.37	35.25	49.25
Avg.	41.67	38.02	48.12	36.27	34.72	49.59

The experiments were performed for the classification of 4000 data images having 40 classes and the operation is done on SVC, BPXNC, KNNC, QDC, LDC and NMC classifiers. The results for cross validation is found better compared to using fixed training data. Also from the Table 2, it is clear that minimum average error rateof k-fold method is obtained for LDC classifier.

IV. CONCLUSION

Compound character is one of the features of Hindi script. This paper presents a system for Hindi script handwritten compound character recognition. In this paper various classifiers have used on a compound character dataset. The dataset was developed using handwriting of various writers. The recognition of compound characters is done by using LDC, QDC, NMC, BPXNC and KNNC classifiers. The gradient features are used in this work. The best result is obtained when a LDC classifier is used. The recognition rate of compound characters can be further increased by increasing the size of the database or may be by using some efficient segmentation algorithm.

REFERENCES

[1] Apurva A. Desai, "Gujarati handwritten numeral optical character recognition through Neural Network", Elsevier, Pattern Recognition 43 (2010) 2582-2589.

[2] Mayuri Rastogi, Sarita Chaudhary, Shiwani Agrawal, "Different Classification Techniques for Character Recognition: A Survey", MIT International Journal of Computer Science & Information Technology Vol 3, pp. 30-34, Jan. 2013.

[3] R. Dineshkumar and J. Suganthi, "Sanskrit Character Recognition System using Neural Network", *Indian Journal of Science and Technology*, Vol 8(1), PP no. -65–69, January 2015.

[4] Rajiv Kumar and Kiran Kumar, "Handwritten Devnagari Digit Recognition: Benchmarking On New Dataset"Journal of Theoretical and Applied Information Technology, Vol. 60 No.3, pp no. 543 – 555, February 2014.

[5] Monica Patel and Shital P. Thakkar, "Handwritten Character Recognition in English: A Survey", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 2,pp no. 345 – 350, February 2015.

[6] Md. Mahbubar Rahman, M. A. H. Akhand, Shahidul Islam and Pintu Chandra Shill, " Bangla Handwritten Character Recognition using Convolutional Neural Network", *I.J. Image, Graphics and Signal Processing*,pp no.- 42-49,Aug 2015.

[7] R. Dineshkumar and J. Suganthi, "Sanskrit Character Recognition System using Neural Network", *Indian Journal of Science and Technology*, Vol 8(1), PP no. -65–69, January 2015.

[8] Te´ofilo E. de Campos, Bodla Rakesh Babu and Manik Varma, "Character Recognition In Natural Images", April 2014.

[9] Amit Choudhary, Rahul Rishi, Savita Ahlawat, "Offline Handwritten Character Recognition using Features Extracted from Binarization Technique", AASRI Procedia 4 pp. 306-312, 2013.

[10] Neha Sahu, R. K. Rathy, Indu Kashyap, "Survey and analysis of Devanagari Character recognition Techniques using Neural Networks", International Journal of Computer Applications Volume 47-No. 15, June 2012.

[11] Paulose Raj, Amitabh Wahi, "Zone based method to classify Isolated Malyalam Handwritten Characters using Hu- Invariant Moments and Neural Networks", IJCA pp. 0975-8887, 'ICIIIOSP-2013'.

[12] R.J. Ramteke, "Invariant Moments based feature extraction for Handwritten Devanagari Vowels Recognition", IJCA, Volume 1-No.18.

[13] Dayashankar Singh, Sanjay Kr. Singh, Dr. Maitreyee Dutta, "Hand Written character recognition using twelve directional Feature Input and Neural Network", IJCA, Volume 1-No.3.

[14] C.Namrata Mahender, K.V.Kale, "Structured based Feature Extraction of Handwritten Marathi word", IJCA, Volume 16-No.6, Feb. 2011.

[15] K.V.Kale, Prapti D. Deshmukh, "Zernike moment based feature extraction for handwritten Devanagari(Marathi) compound character recognition", International Journal of Advanced Research in Artificial Intelligence, Vol.3, No.1, 2014.

[16] Sushama Shelke, "Multistage handwritten Marathi compound character recognition using Neural Networks", Journal of Pattern Recognition Research 2 (253-268), August 2011.

[17] Aamir Khan, Hasan Farooq, "Principal Component Analysis-Linear Discriminant Analysis Feature Extraction for Pattern Recognition", International Journal of Computer Science Issues, Vol. 8, Issue 6, no.2, Nov 2011.

[18] Tian-Fu Gao, Cheng-Lin Liu "High accuracy handwritten character recognition using LDA-based compound distances", ELSEVIER, Pattern Recognition 41 ,pp. 3442-3451, 2008.